2

Naval Submarine Medical Research Laboratory

NSMRL Report 1095 29 November 1988

PSYCHOMETRIC FUNCTION RECONSTRUCTION FROM ADAPTIVE TRACKING PROCEDURES

by

Marjorie R. Leek
Thomas E. Hanna
Lynne Marshall

Naval Medical Research and Development Command
Research Work Unit #M0100.001-5001

Released by:

C. A. HARVEY, CAPT, MC, USN
Commanding Officer
Naval Submarine Medical Research Laboratory

Approved for public rclease; distribution unlimited

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b RESTRICTIVE MARKINGS

2a SECURITY CLASSIFICATION AUTHORITY

2b DECLASSIFICATION / DOWNGRADING SCHEDULE

3 DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release;
distribution unlimited

4 PERFORMING ORGANIZATION REPORT NUMBER(S)
NSMRL REPORT # 1095

5 MONITORING ORGANIZATION REPORT NUMBER(S)
NA

6a NAME OF PERFORMING ORGANIZATION
Naval Submarine Medical
Research Laboratory

6b OFFICE SYMBOL
(If applicable)

7a NAME OF MONITORING ORGANIZATION
Naval Medical Research and Development
Command

6c. ADDRESS *(City, State, and ZIP Code)*
Naval Submarine Base New London
Groton, CT 06349-5900

7b. ADDRESS *(City, State, and ZIP Code)*
NMCNCR, Bethesda, MD 20814-5044

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
NMRDC and ASEE

8b. OFFICE SYMBOL
(If applicable)

9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS *(City, State, and ZIP Code)*
Same as 7B
ASEE, 11 Dupont Circle, Suite 200,
Washington, DC 20036

10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO
65856N	M0100	001	5001

11 TITLE *(Include Security Classification)*
(U) Psychometric function reconstruction from adaptive tracking procedures

12 PERSONAL AUTHOR(S)
Marjorie Leek, Thomas E. Hanna, and Lynne Marshall

13a TYPE OF REPORT
Interim

13b TIME COVERED
FROM ________ TO ____

14. DATE OF REPORT *(Year, Month, Day)*
29 Nov 1988

15. PAGE COUNT
20

16 SUPPLEMENTARY NOTATION

17 COSATI CODES

FIELD	GROUP	SUB-GROUP

18 SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Adaptive testing procedures; Psychometric function; Monte Carlo simulation; threshold and slope parameter estimation

19 ABSTRACT *(Continue on reverse if necessary and identify by block number)*

Adaptive psychophysical procedures have come into widespread use for the estimation of psychophysical performance. Their popularity arises from their speed of implementation and efficiency in that stimulus levels far removed from the selected target values are seldom presented. Thus, experimental time and subject energy can be devoted to a precise delineation of performance at and around the target or threshold region. However, sometimes it is valuable to be able to describe a subject's performance across a wide range of stimulus values by construction of a psychometric function showing how performance changes with changing stimulus values. Since adaptive tracking procedures are specifically assigned to avoid stimulus levels far from the target value, the psychometric function constructed from the data in the track may be precisely defined near the target where there are many trial presentations for each level, but be a poor reflection of performance at levels removed from the target. While some authors have attempted to analyze the trial-by-trial data produced by an

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED □ SAME AS RPT □ DTIC USERS

21 ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a NAME OF RESPONSIBLE INDIVIDUAL
Susan D. Monty, Publications Office

22b TELEPHONE *(Include Area Code)*
(203) 449-3967

22c OFFICE SYMBOL
421

Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19 Cont'd.

adaptive track to construct a psychometric function, there is little evidence that the functions they report do, in fact, represent the underlying function governing subject performance.

A series of computer simulations was undertaken to assess the validity and accuracy of psychometric functions generated from data collected in adaptive tracking procedures. Estimates of both target threshold values and psychometric function slopes were obtained from trial-by-trial data in simulated adaptive tracks and compared with the true values on the functions used to generate the tracks. Simulations were carried out for four psychophysical procedures and two target performance levels, with tracks generated by three different psychometric functions.

The reconstructed psychometric functions generally were accurate reflections of the underlying functions. Threshold estimation was most reliable for most testing methods using the reconstructed functions, rather than calculating mean levels within the track. The two-alternative forced choice procedure produced the poorest slope and threshold estimates.

Psychometric Function Reconstruction from Adaptive Tracking Procedures

by

Marjorie R. Leek
University of Minnesota

and

Thomas Hanna and Lynne Marshall
Naval Submarine Medical Research Laboratory

NSMRL Report No. 1095

Approved and Released by:

C. A. Harvey

C. A. HARVEY, CAPT, MC, USN
Commanding Officer
Naval Submarine Medical Research Laboratory

Approved for public release; distribution unlimited.

SUMMARY PAGE

The Problem

To find an effective psychophysical method for estimating changes in performance as a function of signal level, i.e., for estimating the steepness (slope) of the psychometric function.

The Findings

Computer simulations of adaptive tracking techniques indicate that reconstruction of the psychometric function from the subject's responses at the various signal levels used during the track provided fairly reliable estimates of the slope of the psychometric function. Best estimates were obtained from three-alternative and four-alternative forced-choice procedures. Figures are included that show the effects of psychophysical procedure, number of trials, and slope of the psychometric function on the reliability of estimates of the slope of the psychometric function. Additional figures show the reliability for estimating the subject's "threshold," i.e., signal level required for a criterion level of performance.

Application

The results suggest that current adaptive tracking procedures, in addition to providing estimates of threshold, could provide estimates of the slope of the psychometric functions at no extra cost in terms of experimental time. The figures indicate which procedures would be most efficient, the number of trials required to obtain a given degree of accuracy, and the degree of accuracy that might be achieved.

ADMINISTRATIVE INFORMATION

This research was carried out under Naval Medical Research and Development Command Work Unit 65856N - M0100.001-5001, "Auditory Sonar." The first author was an American Society for Engineering Education summer faculty fellow at NSMRL. This report was submitted for review on 9 April 1987, approved for release on 29 November 1988, and designated as NSMRL Report No. 1095.

Abstract

Adaptive psychophysical procedures have come into widespread use for the estimation of psychophysical performance. Their popularity arises from the speed of implementation and efficiency in that stimulus levels far removed from the selected target values are seldom presented. Thus, experimental time and subject energy can be devoted to a precise delineation of performance at and around the target or threshold region. However, sometimes it is valuable to be able to describe a subject's performance across a wide range of stimulus values by construction of a psychometric function showing how performance changes with changing stimulus values. Since adaptive tracking procedures are specifically designed to avoid stimulus levels far from the target value, the psychometric function constructed from the data in the track may be precisely defined near the target where there are many trial presentations for each level, but be a poor reflection of performance at levels removed from the target. While some authors have attempted to analyze the trial-by-trial data produced by an adaptive track to construct a psychometric function, there is little evidence that the functions they report do, in fact, represent the underlying function governing subject performance.

A series of computer simulations was undertaken to assess the validity and accuracy of psychometric functions generated from data collected in adaptive tracking procedures. Estimates of both target threshold values and psychometric function slopes were obtained from trial-by-trial data in simulated adaptive tracks and compared with the true values on the functions used to generate the tracks. Simulations were carried out for four psychophysical procedures and two target performance levels with tracks generated by three different psychometric functions.

The reconstructed psychometric functions generally were accurate reflections of the underlying functions. Threshold estimation was most reliable for most testing methods using the reconstructed functions, rather than calculating mean levels within the track. The two-alternative forced choice procedure produced the poorest slope and threshold estimates.

I. INTRODUCTION

Adaptive testing prccedures have become popular in psychophysical experiments over the past twenty years due to their efficiency and speed. In these procedures, the level of a stimulus on each experimental trial is determined by performance on the previous trial or trials. Such methods are characterized by their ability to rapidly converge on a given level of performance, and to concentrate the experimental trials in the vicinity of the final measurement of interest. Little experimental time and subject energy is expended on trials placed far from the point of interest on the psychometric function.

The trade for this high efficiency, however, is the loss of information about the underlying form of the function which defines the subject's responses to a hypothetically wide range of stimuli. Since few estimates of performance are obtained at levels removed from the threshold, the function may be well-defined near that point with considerably less precision at the extremes of the function. Although in many studies this price is easily paid, with minimal reduction in important information about subjects' performance, there are instances when more complete descriptions of performance are desirable. This is most notably true when new phenomena are under investigation, or when an interaction between subject variables and stimulus variables is unknown to the extent that performance across stimulus levels cannot be estimated adequately based on one performance level and a review of the pertinent literature. In such a case, adaptive methods may not be the procedure of choice, and speed of experimentation may have to be sacrificed to allow a more complete investigation of the entire psychometric function.

Some experimenters have ignored this problem and have generated psychometric functions based on the listeners' performance on the levels determined by the adaptive track. However, there is little evidence that reliable and unbiased estimates of psychometric function slope can be obtained from a post-hoc analysis of responses. Levitt (1971) discussed the optimal choice of signal levels for estimating either the threshold or slope of a psychometric function and suggested that a reasonable compromise is available to estimate both. Hall (1981) simulated the accuracy and reliability of post-hoc fits to data obtained with a four-interval forced-choice (4AFC) PEST procedure. His results indicate that slope estimates are biased, and he specified the degree of reliability which theoretically can be achieved.

Although little effort has been expended on assessing the quality of slope estimates, much research has been performed on the properties of estimates of thresholds. Numerous factors potentially can influence threshold estimates and might also affect slope estimates. Since one is usually not interested only in slope estimates, it is desirable to choose a procedure which does well in estimating both threshold and slope. It is, then, important in designing a procedure for estimating slope to consider factors which affect threshold estimates, such as psychophysical procedure, target performance level, track length, and actual slope of the underlying psychometric function.

A. Psychophysical Procedures

McKee, Klein & Teller (1985) described the changes in variability to be expected when chance performance levels are altered by the selection of different psychophysical procedures. Whereas measures of performance can extend from a chance level of 0% correct to perfect performance (100% correct) for a free response procedure, that range of performance is halved for a two-alternative forced choice (2AFC) procedure (i.e., chance performance increases to 50% correct). Likewise, changes in chance level for three-alternative (3AFC) (33%) and four-alternative (4AFC) (25%) forced-choice methods alter the range of possible performance. These changes in range of performance levels affect not only the variability of measurement, but are also reflected in the slope of the resulting psychometric function: the slope for a 2AFC-generated function is half that for a function generated in a free response experiment, even though the sensitivity of the listener to the experimental manipulation is, of course, unaltered.

Although several authors have described the statistical properties of 2AFC as less than optimal, it is often used by psychoacousticians due to its criterion-free characteristics (Green and Swets, 1966) and its speed and simplicity of implementation. Rose, Teller & Rendleman (1970) reported computer simulations indicating that the 2AFC procedure produced estimates which were not only biased, but also were characterized by large variability. Kershaw (1985) also demonstrated much greater variability in threshold estimates obtained with the 2AFC procedure than with a yes-no procedure. Kershaw argued that the decision to use a 2AFC procedure must be accompanied by a willingness to present many more trials than would be necessary in a yes-no task, and that even then, rather larger estimation biases are to be expected. Similar conclusions were reached by McKee, et al. (1985). Nonetheless, this method has been used extensively over the past several decades and continues to be attractive because of its saving of experimenter and subject time.

Some experimenters have advocated the use of a three-alternative rather than a two-alternative procedure. Shelton and Scarrow (1984) compared both procedures using two different adaptive methods. While the threshold estimates they obtained were similar, especially for sets of at least 100 trials, the variability was larger for the 2AFC procedure. Hall (1983) also favored a 3AFC procedure, arguing that the lower chance probability produces more stable thresholds and a faster convergence on a threshold value. Tyler, Summerfield, Wood & Fernandes (1982) selected a 3AFC procedure not because of its statistical properties, but because they felt the task was a simpler and more comfortable one for the listeners, allowing the use of a strategy of selecting the "odd" alternative. Most recently, Schlauch and Rose (1986) compared simulations of several psychophysical procedures and found greater variability for the 2AFC than for either the 3- or 4AFC. Interestingly, these authors selected track length based on equal experiment times necessary for the different procedures. They reported that the savings in trial duration for the 2AFC over the multi-interval procedures was not sufficient to make up for the variability in measurement of the former.

B. Choice of Target Level, Length of Track, and Step Size

Adaptive tracking procedures have been described which will allow a wide choice of target performance level depending on the requirements of the experiment and the inclinations of the investigator (Levitt, 1971). Typically, target level is chosen to be some point midway between chance and perfect performance on the psychometric function. Many experimenters have chosen a level of 71% correct as a convenient target level near the mid point (75% correct) on a 2AFC psychometric function. However, the same logic might point to the use of 79% correct, which has the added advantages of being a more statistically stable point (according to McKee et al., 1985). However, it is possible that in an adaptive tracking procedure it might take longer to converge on a 79% level than a 71% level, in that three correct answers are required for a decrease in stimulus level, rather than just two for the 71% level.

The choice of an adaptive track length must be a compromise between the desire for precision of measurement and the need for speed in the experimental procedure. Shelton, Picardi and Green (1982) reported that a 50-trial track obtained with a 2AFC procedure provides reliable threshold estimates. Various investigators have selected track lengths by the number of reversals provided by the track rather than by number of trials in order to assure an adequate sample of the threshold region of te psychometric function. Speed is again the issue here: the shorter the track length, without sacrificing precision, the more efficient the measurement of psychophysical performance.

This paper describes a series of computer simulations undertaken to assess the validity of generating psychometric functions from trials in an adaptive track. Two separate issues are addressed. First, the precision with which such methods indeed do reflect the underlying psychometric function, both in slope and in placement along a stimulus dimension, will be evaluated by comparing the characteristics of the true function (used to generate the adaptive track) with the function reconstructed from the trial-by-trial data in the track. Secondly, the threshold estimate provided by a typical method of locating threshold from an adaptive track, i.e., the mean stimulus level of runs within the track, will be compared with both the true threshold, taken from the known generating function, and the threshold estimate calculated from the reconstructed function.

The accuracy of the reconstructed functions, as well as the reliability of measurement from adaptive tracks will, of course, be affected by the statistical properties of the psychophysical procedure employed and the number of trials evaluated. Therefore, these simulations were performed using four different procedures, with two tracking target levels (71% and 79% correct) and six track lengths.

II. METHOD

Simulations were performed for several adaptive procedures. These procedures varied in the following ways.

a. Psychophysical Procedure. Four different psychophysical procedures were examined: 2AFC, 3AFC, 4AFC, and free response. These procedures produce psychometric functions with differing levels of chance performance, varying from 50% for the 2AFC task to 0% for the free response task.

b. Target Threshold Level. Either 71% or 79% correct level of performance was estimated by using a two-down, one-up rule or a three-down, one-up rule. That is, two (or three) consecutive correct responses led to a decrease of the signal level, and a single incorrect response led to an increase of the signal level (Levitt, 1971).

c. Track Length. Track lengths of 50, 100, 200, 300, 400, and 600 were investigated.

In addition, to simulate different subject performance, three psychometric function slopes were used. Based on results obtained as part of a study reported by Marshall and Jesteadt (1986), these slopes were .2, .4, and .8 z-score units per dB for the free response task. Functions for the other psychophysical procedures were derived from the free-response functions by the

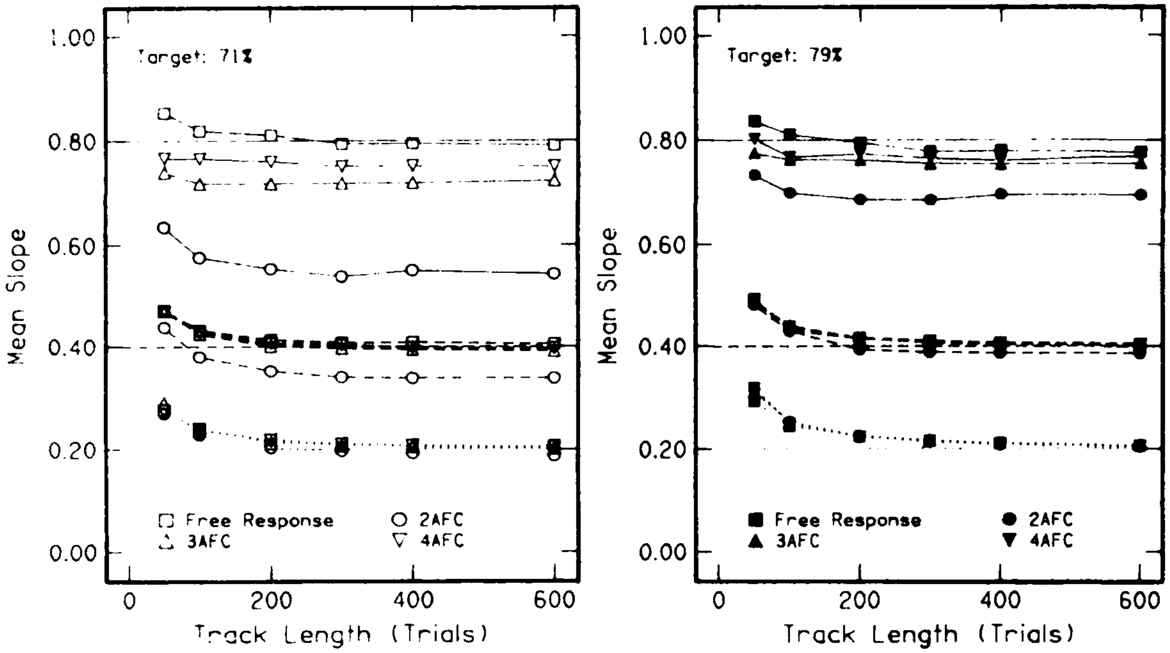

Figure 1. Mean slope estimates from reconstructed functions as a function of length of adaptive track. The left panel and open symbols show results for a target level of 71% correct; the right panel and solid symbols are for a target level of 79% correct. The horizontal lines at slopes of .2, .4, and .8 indicate the underlying slopes of functions generating the tracks, and the results associated with each slope are shown in the same type of line. The parameter in each panel is the psychophysical procedure.

transformation:

$$p_i = 1/i + ((i-1)\ p_{fr}$$

where i is the number of intervals, pfr is the proportion of correct responses in the free response task, and p_i is the proportion of correct responses in the i-interval task.

For each combination of the above variables, 1000 independent simulated tracks were generated. The signal started 6 dB above the midpoint of each function. On each "trial," a random number from 0 to 1 was compared to the probability of detection given the current signal level and psychometric function for that simulation. A random number less than the current probability of detection generated a correct response; otherwise an incorrect response was generated. The level was adjusted in 2-dB steps according to the selected adaptive rule until the specified number of trials was completed.

For each simulated track, a cumulative normal psychometric function was fit to the data using probit analysis (Finney, 1971). The estimates of slope and signal level needed for the target level was estimated by averaging the upper and lower levels of each ascending run after the first two reversals (Levitt 1971). Summary statistics of these three estimates were computed for the 1000 simulations for each condition.

III. RESULTS

A. Slopes of reconstructed functions: Figure 1 shows the mean calculated slopes for the reconstructed functions generated by underlying functions with the three slope values. Track length is shown on the abscissa. The left panel shows results for a target level of 71% correct; the right panel displays slopes for a target of 79%. The horizontal lines indicate the underlying slopes of .2, .4, and .8 z/dB, referenced to the free-response task, and the solid, dashed, and dotted lines identify the data with the appropriate underlying slope. The psychophysical procedures are indicated by different symbols.

Slope estimates stabilize with longer track lengths, although they do not always converge on the underlying slope value. Little improvement in slope estimates is seen for tracks longer than 200 trials. For track lengths less than 200 trials, the reconstructed slope is high relative to the asymptotic value. For tracks longer than 200 trials, the reconstructed slope generally provides a good estimate of the actual slope for underlying slopes of .2 and .4. For the steeper underlying function (.8), the slope estimates generally are lower than the actual slope.

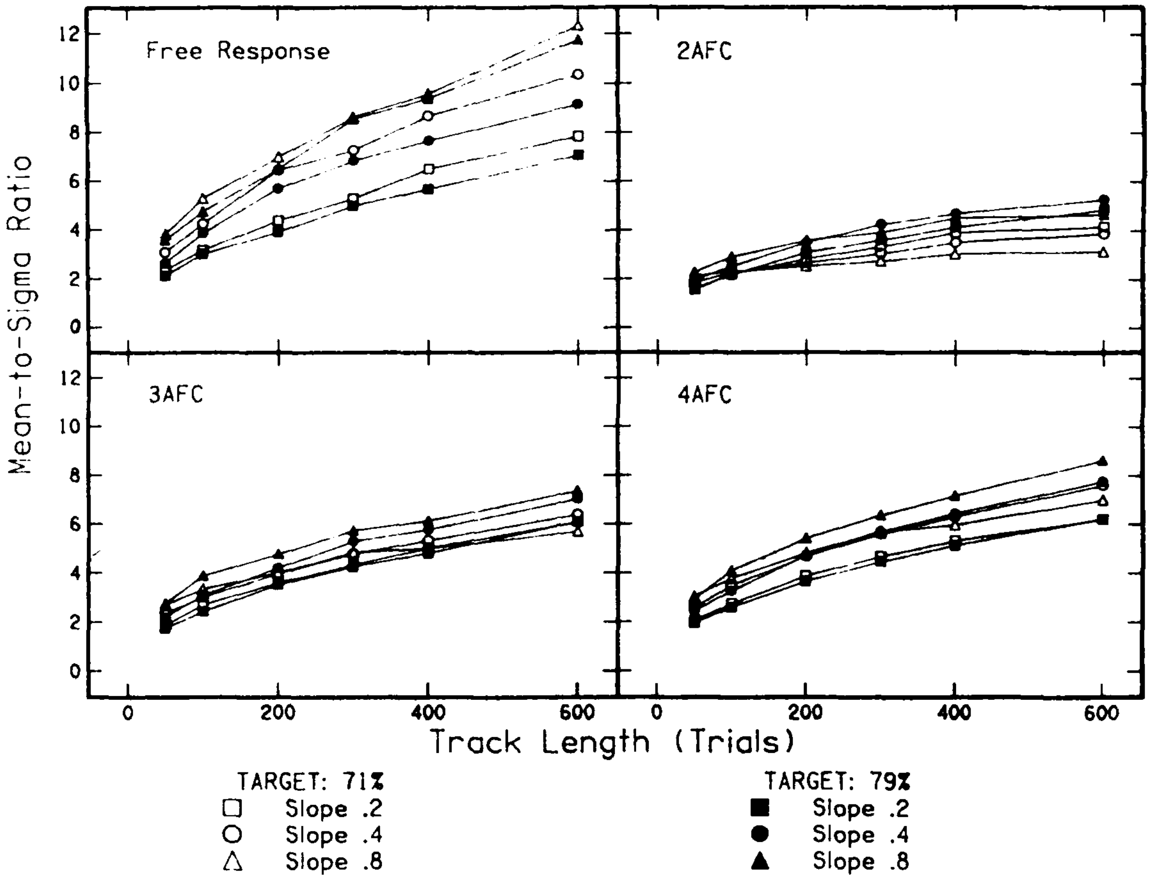

Figure 2. Relative variability of slope estimates for the four procedures. The parameters in each panel are slope and target level.

Differences among procedures are primarily at high slopes. The best slope estimates are consistently provided by the free-response and 4AFC procedures, with the free-response procedure unbiased even at high slopes. The 2AFC procedure produces noticeably poorer slope estimates than any of the other procedures at slopes of .4 at 71% target level and .8 at both target levels.

In order to produce fairly unbiased estimates of slope from the reconstructed functions, track lengths of 100 trials seem adequate for all but the 2AFC procedure. For the 2AFC procedure, the slope estimates often are biased no matter how many trials are used.

Figure 2 displays a measure of variability in slope estimates. The ratio of mean slope to the standard deviations of the slope measurements is shown as a function of track length. This metric indicates the precision of measurement of slopes, independent of the slope value. The larger the mean-to-sigma ratio, the greater the precision of estimate of the slope of the function. The four panels on this figure show the four procedures, with the generating slope and target level as the parameters in each panel. Confidence intervals for the slope estimates can be constructed with these values. For example, for 3AFC with a slope of 0.4, a 200-trial block produces a slope estimate accurate to plus or minus 25%.

The free-response procedure has the greatest precision. Increasing the number of trials on the free-response procedure from 50 to 600 produces a three-fold increase in mean-to-sigma ratio. The 2AFC procedure is less precise than the free-response procedure, with an increase in the number of trials resulting in no more than a doubling of the mean-to-sigma ratio across track length. The 3AFC and 4AFC procedures are at intermediate values.

For a 4AFC PEST procedure, Hall (1981) obtained mean-to-sigma ratios of approximately 1.5 and 3 for track lengths of 50 and 200 trials, respectively. These values are smaller than corresponding values of roughly 2.5 and 4 for the 4AFC procedure examined here. Thus, the adaptive staircase method may provide better estimates of slope than the PEST procedure, probably due to a greater tendence of PEST to concentrate trials on levels nearer threshold, thereby reducing the distribution of testing levels needed to estimate slope.

The trade between reduced variability and length of the track can be shown by the use of the "sweat factor" defined by Taylor and Creelman (1967). This is a measure of the efficiency of a psychophysical procedure based on the variance of threshold estimates resulting from a specified number of trials. The sweat

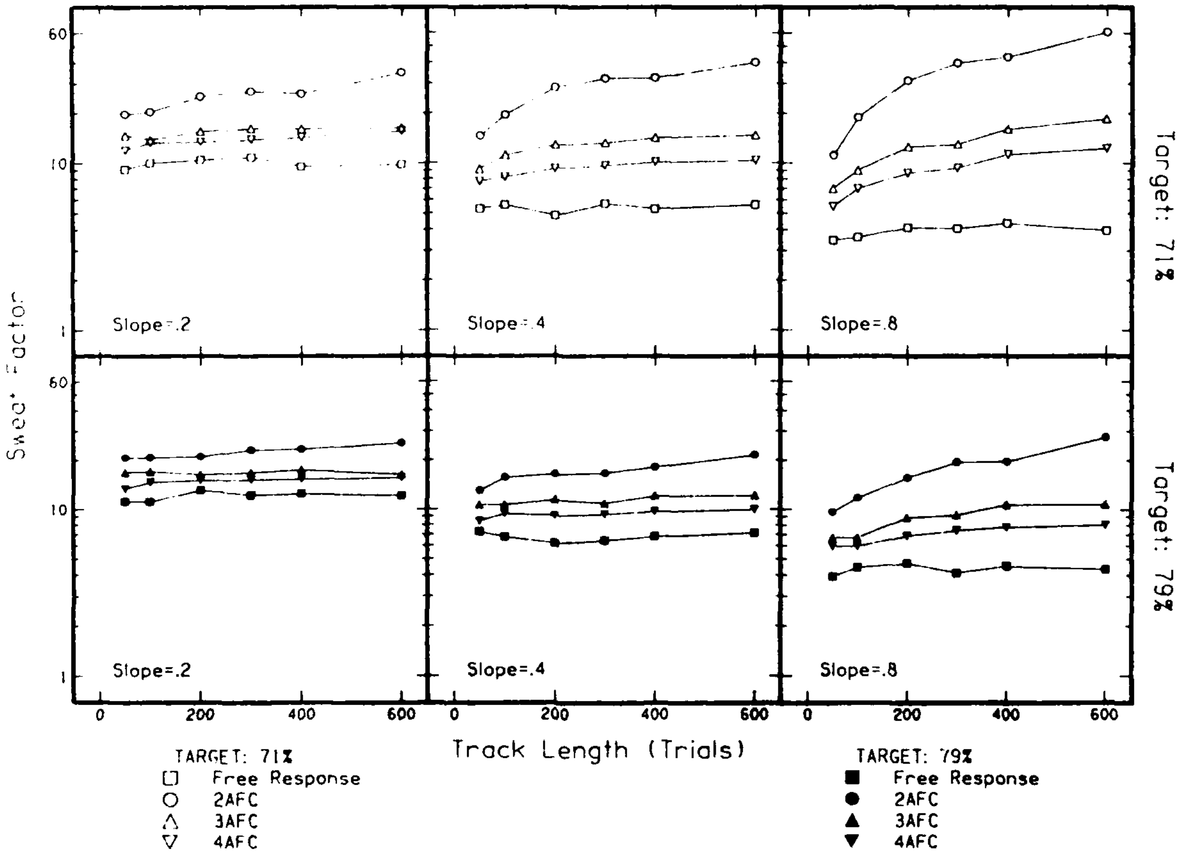

Figure 3. Sweat factors associated with mean-to-sigma ratio as a function of track length. The three columns of panels represent the three slopes. Target levels are shown in the two rows. The parameter in each panel is the psychophysical procedure.

factor, k, is defined as:

$$k = No^2$$

where N is the number of trials in the track, and o2 is the variance of threshold estimates produced. The smaller the value of k, the greater the efficiency of measurement. This metric can be applied to variability of slope estimate by substituting the square of the sigma-to-mean ratio for sigma in the sweat factor equation.

Figure 3 shows sweat factors associated with mean-to-sigma ratios as a function of track length. For the free-response procedure, precision increases with greater number of trials in the track, which results in constant sweat factors across track length. 3AFC and 4AFC also have constant sweat factors across track length, except for 71%, and to a lesser extent, for 79% at the steepest slope value (.8). For the 2AFC procedure, the sweat factor increases with increasing number of trials. Because there is so little increase in precision with number of trials, it is inefficient to use more trials.

For all procedures, there is greater precision (and thus greater efficiency) at higher slopes. Since a steep slope may also be interpreted as a large step size with a shallower psychometric function, larger step sizes result in greater precision 1. This is particularly true for the free-response procedure at longer track lengths.

B. Threshold Estimation. Figure 4 shows the mean error in threshold estimation for each of the procedural conditions studied as a function of track length. These values were calculated by subtracting the threshold value at the targeted level on the underlying psychometric function from the mean threshold estimates obtained from the simulations. The top row of panels show mean error using the midrun means of the track; the bottom row displays the error associated with determining threshold from the reconstructed psychometric function. From left to right the panels represent an underlying slope of .2, .4, and .8 z-units per dB, respectively.

Most of the mean error values fall within 1 dB of the true threshold value. In almost all cases, the threshold value calculated from the reconstructed psychometric function is closer to the true value than is the midrun mean estimate, which tends to overestimate the true threshold value. Estimates over all condi ions are generally stable with increases in track length after about 100 trials. With the exception of 2AFC-71%, no clear advantage in threshold estimation may be seen for either of the two targeted threshold values (71% and 79%). The 2AFC-71% procedure is significantly worse than the other procedures. The mean error was not stable by 100 trials -- in fact, for the threshold estimates based on the midrun means, the error actually increased with increasing track length for all slopes.

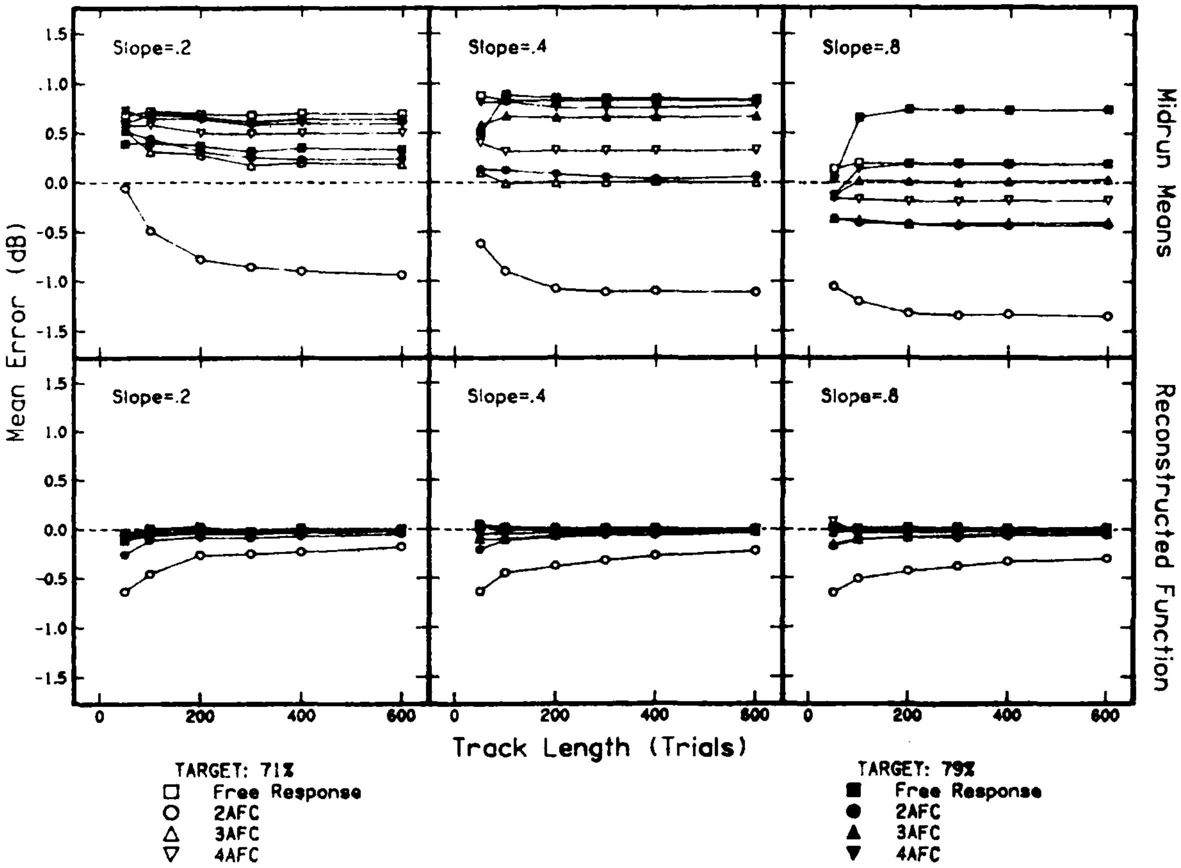

Figure 4. Mean error in threshold estimate as a function of track length. The top row of panels shows errors based on midrun mean calculations of threshold; the bottom row is based on threshold calculations from the reconstructed psychometric function. Open symbols reflect a target level of 71%; solid symbols indicate a target level of 79%. Different symbols indicte the psychophysical procedure used.

Figure 5 displays the variability in the threshold measurements for the various conditions as the mean standard deviation of threshold estimates over the 1000 simulations for each condition as a function of track length. The panels and parameters are the same as for the earlier figure showing threshold error. As expected, all conditions show a decrease in variability with increases in track length. The reconstructed-function thresholds at the target level of 79% generally show the smallest standard deviations. The midrun mean estimates almost always result in greater variability than do the estimates from the reconstructed functions. Once again, the 2AFC-71% procedure is worse than all the others, especially for threshold estimates from reconstructed functions.

Figure 6 shows the sweat factors associated with the various procedures and target levels simulated here as a function of track length. Note that in some conditions there is a slight decrease in the sweat factor with increasing track length from 50 to 100 trials, but little advantage to longer tracks. The 2AFC condition is the least efficient measure of all procedures and slope values; the free response is the most efficient. A small advantage is observed for the reconstructed function threshold determinations over the midrun-mean estimations. Although the 79% level is generally more efficient, neither 71% nor 79% is consistently better than the other, and the differences are quite small, except for 2AFC where 71% is clearly inferior.

IV. DISCUSSION

A. Track Length. The reliability of measurement is influenced by the number of trials in the adaptive track. Results of the present simulations showed a slight improvement in efficiency between 50 and 100 trials. However, in actual practice, this effect is not seen. Shelton et al. (1982) investigated the precision of threshold measurement for detection of sinusoids in noise, and concluded that minimum track lengths should be about 50 trials for an adaptive staircase procedure such as that simulated here.

Slope estimates showed an increase in stability for tracks of 100 versus 50 trials for all experimental conditions simulated here, although the bias in slope estimates in many cases did not disappear. This suggests that if slope estimates are to be obtained in addition to thresholds, 100-trial tracks are desirable.

B. Target Level. While it has, in the past, been more common to use the 71% level, recently some experimenters have selected a target level higher on the psychometric function. The rationale for this change is partially subjective; subjects may be more relaxed and less anxious if the feedback in an experiment

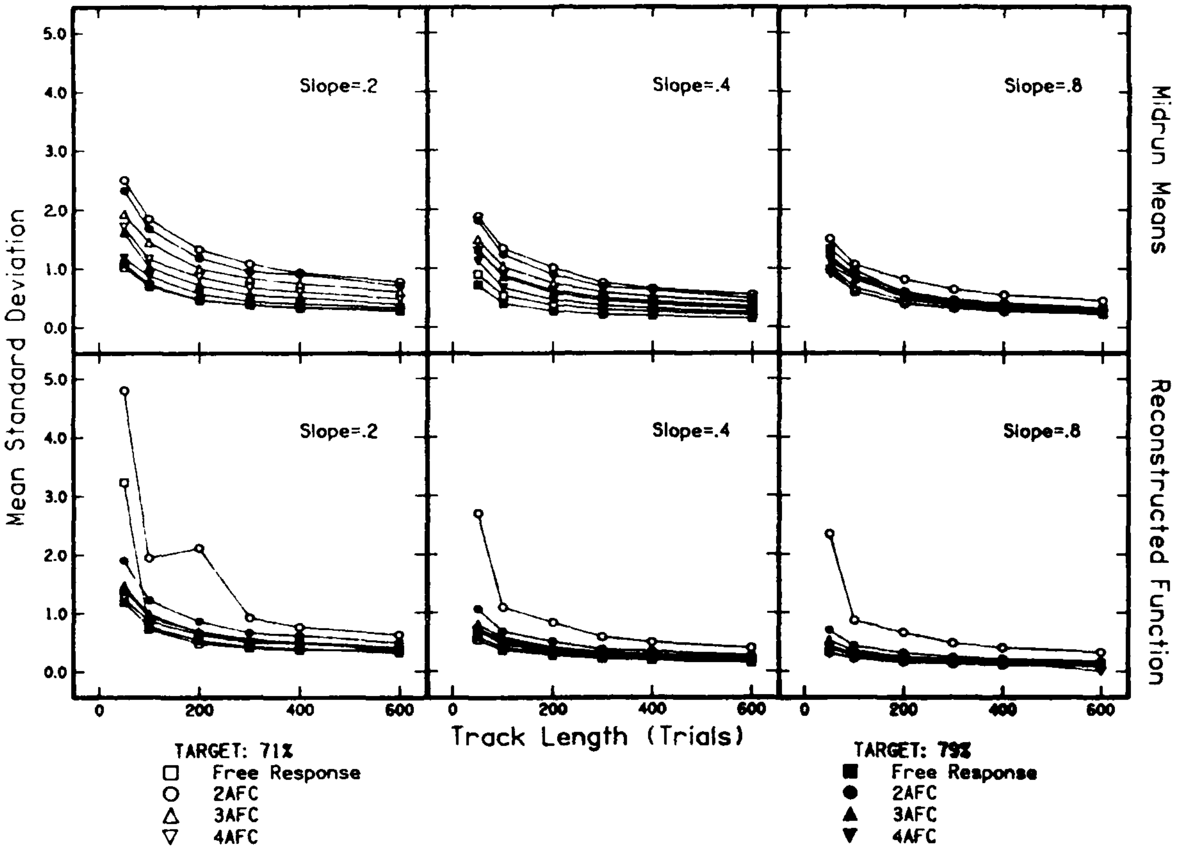

Figure 5. Mean standard deviation of threshold estimates as a function of track length. The panels and symbols are the same as for Figure 4.

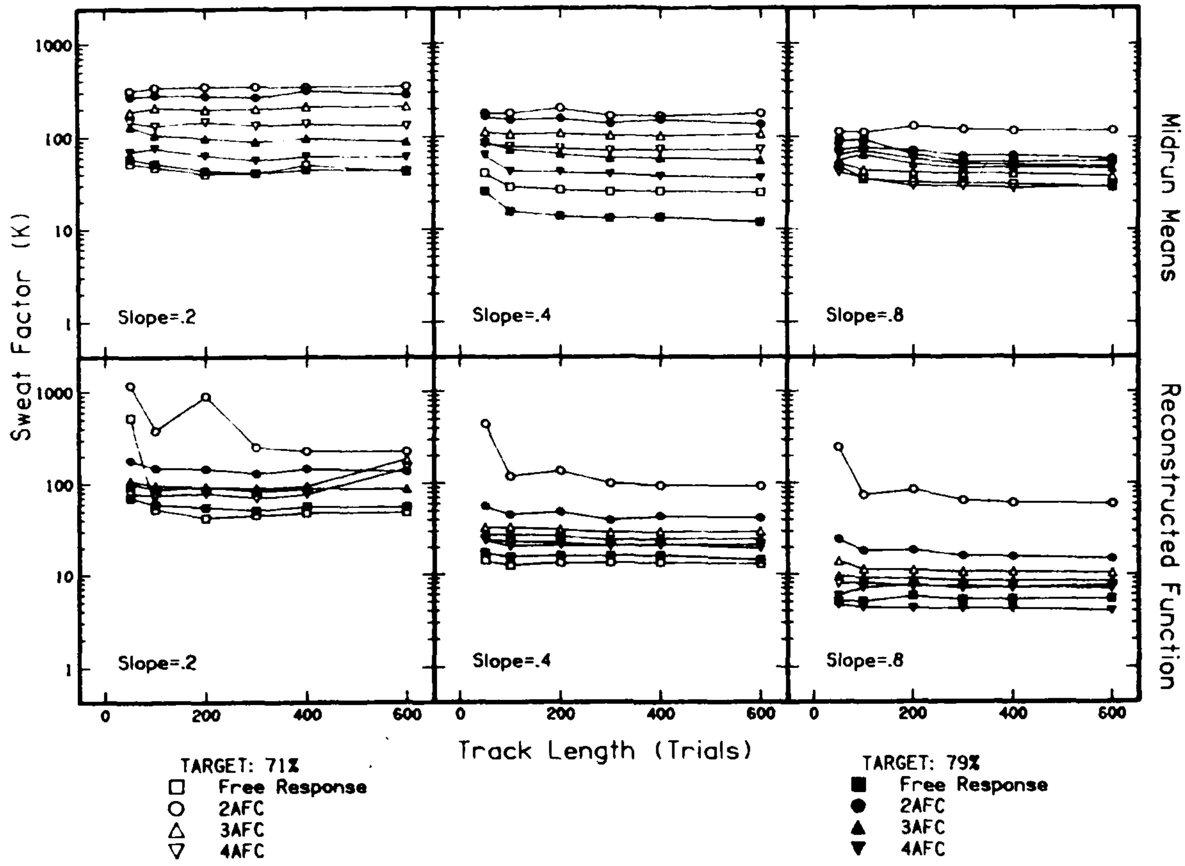

Figure 6. Sweat factor associated with threshold estimates as a function of track length. Panels and symbols are the same as for Figure 4.

notifies them that their performance is more good than bad. Thus, choosing a level of 79% on the function allows the subject to have more "successes," and may result in longer and perhaps more generally satisfactory performance, both from the subject's and the experimenter's point of view.

There are also theoretical reasons why the higher target level should provide more stable estimates of performance, and the smaller threshold variability is reflected in these simulations. McKee et al. (1985) described the asymmetry of variance for psychometric functions generated with 2AFC psychophysical tasks. Because of the compressed function ranging from chance at 50% to 100%, performance on trials placed below the center of the distribution (i.e., below the 75% level) demonstrates greater variability than does performance at levels above the center. This would predict a smaller variance for the 79% target level in these simulations than for the 71% level, especially for the 2AFC procedure. Smaller asymmetries would be expected for the other forced-choice procedures, since both target levels are above the midpoints on these compressed functions.

The higher target level did produce slightly more accurate and reliable estimations of threshold for most of the procedural conditions examined in these simulations. The difference due to target level was, however, most obvious for the 2AFC procedure.

Figure 2 indicates that the effect of target level on variability of slope estimates is inconsequential compared to the more important aspects of choice of psychophysical procedure and steepness of the underlying function, except for the 2AFC procedure. This is not unexpected since the slope measurements are derived from many points on the function rather than just one with an associated variability.

C. Psychophysical Procedure. A more significant problem, however, for estimates of both parameters of the psychometric function is the choice of psychophysical procedure. The somewhat discouraging finding is that the procedure which may be the best from the point of view of limited criterion effects and response bias (the 2AFC) may be the worst choice for precision of measurement because of its psychometric properties. The analyses of the present 2AFC simulations are in agreement with Rose et al. (1970) and Kershaw (1985), who both warned of relatively large estimation biases using the 2AFC procedure, and suggested that more trials (longer track length) might be necessary if this procedure is selected over some other. However, the slope estimates shown in Figure 1 hold little hope for significant improvements in accuracy for the 2AFC procedure, even with longer track lengths. If adaptive tracks are to be analyzed to extract information about the slope of the psychometric function, there is good reason not to use the 2AFC procedure, especially if the underlying slopes are expected to be at all steep, or if a large step size is used.

While the results of these simulations, as well as the reports of other authors, would strongly suggest the use of 3- or 4AFC procedures over the commonly used 2AFC method, many experimenters are reluctant to make the change. The most significant objections to the more psychometrically appropriate procedures are the increased experimental time due to longer trial durations and the nontrivial issue of response bias in the multi-alternative procedures. Johnson, Watson and Kelly (1984) reported significant differences in performance on the individual intervals of a three-alternative forced choice task. Performance tended to be best in the third interval and worst in the first interval. In an appendix to their article, these authors reported some evidence indicating that sensitivity to signals is not different in each interval, but that instead the effect of interval resulted from more central factors such as attention and memory. However, this work does call into question the commonly accepted assumption of equal performance on all intervals of a procedure as long as the probabilities of signal presentation are equal across intervals.

D. Fitting the Psychometric Function: Some caution must be urged in using a probit fit to functions whose range does not extend from 0 to 100% correct. The probit transforms data obtained from forced-choice procedures so that chance performance is treated as 0% correct. In a true 0-100% function, data around the lower boundary have small error. However, in a forced-choice procedure, the lower boundary is associated with relatively large variability. The probit transforms the lower boundary to zero, but makes no transformation of the variability associated with that point. Since the probit analysis weights are also influenced by the variability at each level, this transformation may artificially alter the properties of the forced-choice data. Moreover, this form of bias due to the fitting procedure would be greatest for the 2AFC procedure because the transformed data would have the greatest variability near the lower boundary. The fitting problem would also be greater for the 71% condition because there would be more trials near the lower boundary. In fact, the 2AFC 71% condition was the poorest of the procedures examined here, perhaps due in part to the probit fit. An alternative fitting procedure might improve slope estimates.

E. "Free" Slope Estimates from Adaptive Tracks: Adaptive tracking procedures have been developed to quickly and efficiently obtain accurate performance measures at a targeted point on a psychometric function, with secondary concern for other characteristics of the function. However, these simulations have shown that with a thoughtful choice of procedure, and sufficient trials in the track, reliable estimates of slope of the function can be obtained at the same time. The most accurate and reliable slope estimates resulted from either a free-response or a 3- or 4AFC procedure, with 100 or more trials in the track. Moreover, one can interpret the effect of slope shown in Figure 3 to indicate that a judicious choice of step size could improve the reliability of slope estimates.

V. CONCLUSIONS

One would hardly expect shockingly poor performance with any of the procedures examined here for either the slope or threshold measurements. These procedures have been used extensively by experimenters in psychophysics over a period of many years, and one would expect that ill-behaved procedures would have been dropped from scientists' repertoires. The simulations reported here must be viewed relatively: some procedures are consistently "better" than others in terms of their statistical behavior. There are always other factors of psychophysical performance which must be taken into consideration when planning an experiment which may indicate a procedure which is not optimal as determined by simulated data such as these. This is undoubtedly why, even after years of warnings about the relatively poor psychometric performance of 2AFC procedures, these methods are still in common use in psychophysical laboratories.

It is important, however, that experimenters recognize the limitations of their methodologies and understand the psychometric implications of their choices in designing their experiments. It is for this reason that these simulations were undertaken.

Under the assumptions of these simulations, the generation of psychometric functions from the trial-by-trial data produced by adaptive tracking procedures may be quite accurate. The caveat here, however, is that the reconstructed function was constrained to have the same shape as the underlying function. If the general form of the function is truly unknown, and the assumption of the Gaussian distribution is not valid, the properties of the reconstructed function may miss the mark badly. Fortunately, for most experimentation in psychophysics, the assumption of a cumulative normal function is reasonable, so that the reconstructed functions should accurately reflect perception.

Some improvement in threshold estimation using the 2AFC procedure can be observed if the reconstructed function is used for the calculation, rather than using the mean values of the ascending runs in the track. In fact, most threshold estimates were better in terms of accuracy and variability when determined from the function rather than from the midrun mean calculation.

In summary, performance estimates based on reconstructed psychometric functions should be quite accurate. There is some psychometric advantage to the selection of a 79% target level over 71% for decreased variability of threshold estimates, and, for 2AFC, also for slope estimates. For all conditions and parameters measured, the free-response procedure consistently provided the most reliable results. If a forced-choice method is desired for increased precision of measurement as well as for accurate reconstruction of underlying psychometric function, both the 3- and 4AFC outperform the 2AFC from a psychometric standpoint.

This work was supported by the U.S. Navy and the American Society for Engineering Education. It was undertaken while the senior author was an ASEE Summer Faculty Fellow at the Naval Submarine Medical Research Laboratory, Groton, CT.

FOOTNOTE

[1] Obviously, at some point, increasingly larger step sizes begin to have a detrimental effect, as discussed by Levitt, 1971. This holds true for all the results in this paper where improvements in test performance occur with larger step sizes.

REFERENCES

Finney, D. J. (1971). Probit Analysis. Cambridge: The University Press.

Green, D. M., and Swets, J. A. (1966). Signal Detection Theory and Psychophysics. Huntington, NY: Krieger.

Hall, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. Journal of the Acoustical Society of America, 69, 1763-1769.

Hall, J. L. (1983). A procedure for detecting variability of psychophysical thresholds. Journal of the Acoustical Society of America, 73, 663-667.

Johnson, D. M., Watson, C. S., and Kelly, W. J. (1984). Performance differences among the intervals in forced-choice tasks. Perception & Psychophysics, 35, 553-557.

Kershaw, C. D. (1985). Statistical properties of staircase estimates from two interval forced choice experiments. British Journal of Mathematical and Statistical Psychology, 38, 35-43.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. Journal of the Acoustical Society of America, 49, 467-477.

Marshall, L., and Jesteadt, W. (1986). Comparison of pure-tone audibility thresholds obtained with audiological and two-interval forced-choice procedures. Journal of Speech and Hearing Research, 29, 82-91.

McKee, S. P., Klein, S. A., and Teller, D. A. (1985). Statistical properties of forced-choice psychometric functions: Implications of probit analysis. Perception & Psychophysics, 37, 286-298.

Rose, R. M., Teller, D. Y., and Rendleman, P. (1970). Statistical properties of staircase estimates. Perception & Psychophysics, 8, 199-204.

Schlauch, R. S., and Rose, R. M. (1986). A comparison of two-, three-, and four-alternative forced-choice staircase procedures. Journal of the Acoustical Society of America, 80, S123.

Shelton, B. R., Picardi, M. C., and Green, D. M. (1982). Comparison of three adaptive psychophysical procedures. Journal of the Acoustical Society of America, 71, 1527-1533.

Shelton, B. R., and Scarrow, I. (1984). Two-alternative versus three-alternative procedures for threshold estimation. Perception & Psychophysics, 35, 385-392.

Taylor, M. M., and Creelman, C. D. (1967). PEST: Efficient estimates on probability functions. Journal of the Acoustical Society of America, 41, 782-787.

Tyler, R. S., Summerfield, Q., Wood, E. J., and Fernandes, M. A. (1982). Psychoacoustic and phonetic temporal processing in normal and hearing-impaired listeners. Journal of the Acoustical Society of America, 72, 740-752.